PERSONAGRAM: A MULTIMODAL FRAMEWORK FOR PERSONALIZED EMOTION-AWARE SUBTITLES

Seungkyu Oh[1], *Gwangsoo Kim*[1], *Wookey Lee*[1]

[1] Department of Industrial Engineering, Inha University

ABSTRACT

Subtitles support comprehension but often fall short in conveying the speaker's emotional nuances or adapting to user contexts. This study proposes a multilayered framework that integrates multimodal emotion recognition with personalized subtitle generation. Speech features extracted by Wav2Vec2 and text features extracted by BERT are fused through attention-based mechanisms to achieve fine-grained recognition. Emotions are represented both as categorical classes and in Valence–Arousal–Dominance dimensions, then mapped to color and style attributes. A personalization layer incorporates user profiles such as age and cultural background, allowing identical emotions to be rendered differently according to the viewer. On the IEMOCAP six-emotion classification task, the proposed model surpassed recent state-of-the-art methods. Ablation studies and user evaluations further validated improvements in both recognition accuracy and user experience. Overall, this work moves beyond simple information transfer, demonstrating the potential of combining emotion and personalization for adaptive subtitle generation.

Index Terms— Speech emotion recognition, Personalize, Wav2vec2, Subtitle

1. INTRODUCTION

Subtitles in multimedia content have become a crucial tool for enhancing viewer comprehension. However, most subtitles fail to sufficiently reflect the speaker's emotional nuances or user-specific contexts. Prior research has progressed along three main directions: (1) emotion visualization that expresses emotions through color or font styles, (2) user-centered personalization that adjusts subtitle size and readability, and (3) personalized emotion recognition that improves the precision of speech- and text-based emotion models. Yet, these approaches have remained fragmented, without being integrated into a unified framework that simultaneously addresses emotion recognition and user experience.

In particular, conventional subtitles often merely translate or display scripts, preserving minimal semantic content while neglecting both the depth of the speaker's emotions and interpretations tailored to individual viewers. This limitation prevents viewers from fully empathizing with the context and emotions of real conversations. Thus, emotion-aware subtitles should go beyond simple translation, providing personalized expressions that incorporate contextual factors such as cultural background and age.

To address these limitations, this study proposes a multilayered framework that integrates multimodal emotion recognition with personalized subtitle generation. The proposed model jointly encodes speech and text using Wav2Vec2 and BERT, and leverages cross-attention fusion for fine-grained emotion recognition. The recognized emotions are represented both as categorical classes and as continuous dimensions (VAD), which are mapped to colors (HSL space) and stylistic attributes. Finally, a personalization layer embeds user profiles to transform emotional representations into personalized subtitles. The main contributions of this study are as follows:

1. A novel multilayered framework is introduced that unifies three existing research axes: emotion visualization, user-centered personalization, and personalized emotion recognition.
2. An attention-based multimodal emotion recognition model is implemented, combining speech and text to enable precise emotional representations.
3. Recognized emotions are further transformed into personalized subtitle expressions optimized for cultural background and age, thereby providing experiences that extend beyond simple information transfer.

2. RELATED WORK

Research on enhancing subtitles—whether by improving emotional expressiveness or tailoring to user experience—has advanced along three main streams: (1) personalization of emotion recognition models, (2) visual reflection of speaker emotions in subtitle style, and (3) adaptation of subtitles to user preferences or contexts.

Recent studies on personalized emotion recognition (e.g., Tran et al. (2023)[1], Kargarandehkordi et al. (2024)[2], Li

Thanks to XYZ agency for funding.

et al. (2024)[3]) modeled speaker-specific traits such as pitch and intonation, improving classification accuracy but leaving open how emotions should be expressed to viewers. To address this, researchers explored emotion visualization in subtitles. Early work by Ohene-Djan et al. (2006)[4] used color, size, and animation for hearing-impaired audiences, followed by refinements linking emotions to stylistic features such as color and font [5, 6]. More recent approaches applied large language models [7] or augmented reality [8], though most assumed a standard viewer and overlooked cultural or individual differences.

Another line of research treated subtitles as interactive media. Wu et al. (2025)[9] showed benefits of dynamic positioning, Gorman et al. (2021)[10] enabled adaptive control of style, and Korbar et al. (2024)[11] generated speaker-specific subtitles. While these improved usability, they did not optimize emotional expression.

In contrast, our study integrates these three directions into a unified framework that both recognizes fine-grained speaker emotions and adapts their expression to viewers' cultural background and preferences. This dual-level personalization distinguishes our work from prior approaches.

3. METHODOLOGY

This study proposes a multilayered framework that recognizes emotions from multimodal inputs (speech and text) and transforms them into personalized subtitle expressions. The overall architecture, illustrated in Fig. 1, consists of three main components: (1) multimodal encoding, (2) emotion representation mapping, and (3) a personalization layer.

3.1. Multimodal Encoding

Input speech is processed by a Wav2Vec2 encoder, which produces speech embeddings F_{Audio} while preserving temporal characteristics. In parallel, Whisper transcribes the speech into text, which is then passed to a BERT encoder to obtain text embeddings F_{Text}.

The two modalities are concatenated as:

$$X = [F_{\text{Audio}}; F_{\text{Text}}] \in \mathbf{R}^{B\times(N_s+N_t)\times D} \tag{1}$$

where B is the batch size, N_s and N_t denote the sequence lengths of speech and text tokens, and D is the embedding dimension.

The multimodal embeddings are processed by a Transformer encoder that applies both self-attention and cross-attention, yielding a fused representation Y. Mean pooling is applied to obtain a summary vector:

$$\bar{h} \in \mathbf{R}^{B\times D} \tag{2}$$

3.2. Emotion Representation

The pooled representation $\bar{h}$ is projected into two branches:

Emotion	Valence(V)	Arousal(A)	Dominance(D)
Joy(Y)	0.960	0.648	0.588
Anger(R)	–0.666	0.730	0.314
Sadness(B)	–0.896	–0.424	–0.672
Fear(P)	–0.854	0.680	–0.414
Surprise(O)	0.750	0.750	0.124
Disgust(B)	–0.896	0.550	–0.366
Neutral(W)	–0.062	–0.632	–0.286

Table 1. The 7 core emotions based on Ekman's basic emotions and their corresponding VAD values extracted from a VAD lexicon.

3.2.1. *Categorical emotion classification head*

$$z = W_c\bar{h} + b_c, \quad P = \text{softmax}(z) \tag{3}$$

where P denotes the probability distribution over discrete emotion classes.

3.2.2. *Continuous emotion dimension head*

$$E = (V, A, D) = f(X; W) \tag{4}$$

where E represents Valence, Arousal, and Dominance.

The predicted emotion vector E is transformed into the HSL (Hue, Saturation, Lightness) color space following predefined rules Table 1. Valence corresponds to Hue, Arousal to Saturation, and Dominance to Lightness, thereby determining the chromatic properties of subtitles.

3.3. Personalization Layer

The core novelty of this study lies in extending emotion-aware subtitles into personalized expressions. A user profile is defined as:

$$p_u = \{\text{age}, \text{culture}, \text{gender}\} \tag{5}$$

which is embedded by a Profile Encoder ϕ to produce:

$$u = \phi(p_u) \tag{6}$$

This embedding u is incorporated in two ways:

The base mapping $E = (V, A, D)$ is first projected into HSL. The embedding u then modulates the HSL attributes via Feature-wise Linear Modulation (FiLM):

$$\begin{aligned} \widehat{Hue} &= \gamma_H(u)\, Hue(E) + \beta_H(u) \\ \widehat{Sat} &= \gamma_S(u)\, Sat(E) + \beta_S(u) \\ \widehat{Light} &= \gamma_L(u)\, Light(E) + \beta_L(u) \end{aligned} \tag{7}$$

Culture embeddings capture universal patterns with local variations (Jonauskaite et al., 2020)[12], while gender embeddings reflect differences in saturation, contrast, font weight, and stylistic preferences for curved versus angular shapes. Prior studies report consistent sex differences in color preference across cultures (Hurlbert & Ling, 2007)[13], which

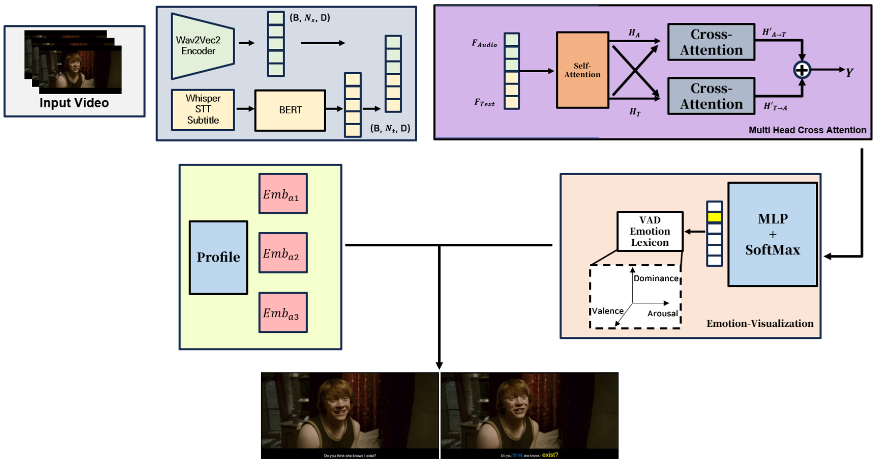

Fig. 1. Overall architecture of the Personagram framework. The input video is processed through Wav2Vec2 and BERT encoders to extract audio and text features, which are fused via self- and cross-attention mechanisms. A user profile embedding is integrated to adapt subtitle styles based on individual preferences. The fused representations are mapped to the VAD (Valence–Arousal–Dominance) emotion space and visualized through emotion-aware coloring. The final output is personalized subtitles that reflect both speaker emotion and viewer profile.

can be explained by the Ecological Valence Theory(Palmer & Schloss, 2010; Al-Rasheed, 2015; 2022) [14] [15] [16]. Furthermore, linguistic and cultural evidence demonstrates that color terms are often associated with gendered meanings (Jonauskaite et al., 2021)[17]. Building on this evidence, we design gender embeddings to finely adjust color (Hue, Saturation, Lightness) and typographic rendering.

In addition to FiLM modulation, rendering strategies are directly adapted based on age. Subtitle attributes such as font size, contrast, simplification level, and the use of emojis or slang are tailored for different age groups. For example, children receive large fonts with intuitive emojis, teenagers encounter slang- or emoji-rich subtitles, and elderly users are provided with high-contrast and readability-focused styles.

Age, culture, and gender are represented as continuous or categorical attributes, respectively, and embedded such that similar values (e.g., adjacent age groups, geographically close cultures, or the same gender group) are placed close together in the embedding space. This facilitates effective FiLM-based modulation and rendering policy learning.

4. EXPERIMENTS

We evaluated the proposed Self- and Cross-Attention-based multimodal emotion recognition model on the IEMOCAP dataset. IEMOCAP consists of approximately 12 hours of conversational data, including both scripted and improvised dialogues performed by professional actors. In this study, audio and transcribed text were used to classify six emotion categories.

Table 2 compares our model with recent state-of-the-art approaches on IEMOCAP. The proposed model achieved WA 76.03% and WF1 76.01%, surpassing prior methods. In particular, unimodal approaches (Wang et al., 2025; Zou et al., 2022)[18] remained around 73–74% WA/WF1, while our multimodal fusion clearly outperformed them. Compared with cross-attention-based fusion (Zhao et al., 2022), our method also showed stable superiority, indicating that the model learned effective bidirectional interactions beyond simple modality concatenation.

Graph-based approaches (e.g., GASMER, GraphSmile) [19] and LLM-based methods (e.g., SpeechCueLLM) [20] reported WF1 scores of 71–73%, which were relatively lower. These results suggest that while graph structures or prompt learning can be useful, they may face limitations if fine-grained multimodal interactions are not captured. By contrast, our Self- and Cross-Attention fusion consistently leveraged complementary information across modalities to achieve high performance.

We additionally conducted a user study to evaluate the subjective quality of the proposed *personalized emotion subtitles*. Participants compared *regular subtitles* against personalized subtitles generated by our model. Evaluation metrics included Clarity, Comprehension, Effort, Readability, Emotion Reflection, Color Accuracy, Engagement, Visual Appeal, Satisfaction, and Recommendation (all rated on a 1–5 Likert scale). Table 3 shows that personalized subtitles consistently outperformed regular subtitles across all dimensions. Notably, improvements were most pronounced for Comprehension (+1.0), Emotion Reflection (+1.0), Color Accuracy (+1.0), and Satisfaction (+0.8), highlighting that emotional expressiveness and accurate color mapping directly enhanced user experience. Furthermore, age-stratified analy-

Table 2. Comparison with recent SER models on IEMOCAP (6-class). Our method achieves state-of-the-art performance in both WA and WF1.

Model	Year	Key Architecture	WA (%)	WF1 (%)
GASMER	2025	GNN, Adaptive Structure Learning	–	71.20
SpeechCueLLM	2024	LLM-Prompt (LLaMA-2)	–	72.60
GraphSmile	2024	GNN, RoBERTa	–	72.77
SSEAN	2025	Dual Recurrent Networks, Commonsense Knowledge	–	73.94
SDT	2023	Transformer, Self-Distillation	74.08	–
Ours	–	Self & Cross-Attention Fusion	**76.03**	**76.01**

Metric	Regular	Personalized	Δ (Gen–Reg)
Clarity	3.8	4.6	+0.8
Comprehension	3.4	4.4	+1.0
Effort	3.4	4.0	+0.6
Readability	3.8	4.4	+0.6
Emotion Reflection	2.3	3.3	+1.0
Color Accuracy	2.7	3.7	+1.0
Visual Appeal	3.1	3.9	+0.7
Satisfaction	3.6	4.4	+0.8
Recommendation	3.5	4.2	+0.8

Table 3. User evaluation scores (1–5) comparing regular and personalized subtitles.

Metric	10s	20s	30s	60s
Clarity	3.63	3.80	3.84	**3.94**
Readability	4.47	4.59	4.56	**4.69**
Emotion Reflection	3.47	3.42	3.40	**3.56**
Visual Appeal	3.53	3.83	4.01	**4.06**
Satisfaction	4.29	4.39	4.42	**4.44**
Recommendation	4.16	4.24	4.22	**4.25**

Table 4. Age-group evaluation (1–5) of personalized subtitles.

sis (Table 4) confirmed consistent benefits across all groups. Teenagers and young adults particularly valued improvements in Engagement and Visual Appeal, while participants in their 30s showed stronger gains in Emotion Reflection and Visual Appeal. Seniors (60+) exhibited the largest improvements in Clarity, Readability, and Satisfaction. These results demonstrate that personalized subtitles are perceived more positively than regular subtitles across all age groups, with especially strong benefits for older viewers.

An ablation study (Table 5) further confirmed the contribution of each modality and the effectiveness of the proposed fusion. Text-only input achieved 67.0% accuracy, audio-only input reached 69.0%, and simple concatenation stayed at 69.0%. In contrast, our full model achieved 76.0%, demonstrating that Self- and Cross-Attention fusion effectively captured complementary multimodal information.

Setting	WA (%)	WF1 (%)
Text-only	67.0	66.5
Audio-only	69.0	68.7
Simple concatenation	69.0	68.9
Ours (Self+Cross)	**76.0**	**76.0**

Table 5. Ablation study on IEMOCAP. Self- and cross-attention fusion yields the largest gains.

5. CONCLUSION

This paper proposed a novel framework that integrates three major directions in emotion subtitle research: (1) emotion visualization, (2) user-centered personalization, and (3) personalization of emotion recognition models. The proposed approach not only generates subtitles based on fine-grained speaker emotion recognition but also adapts their presentation to viewers' cultural backgrounds and preferences, enabling personalized emotion delivery.

Experiments demonstrated that our model outperformed recent state-of-the-art methods on the IEMOCAP dataset. The ablation study confirmed the effectiveness of cross-attention fusion, showing clear improvements over unimodal inputs and simple concatenation. In addition, the user study revealed gains in emotion reflection, color accuracy, and visual satisfaction, validating the proposed method from a user experience perspective. Future work will focus on extending to video modalities, advancing user profiling, and conducting large-scale user studies to further strengthen personalization.

6. REFERENCES

[1] A. Tran et al., “Personalized speech emotion recognition with speaker adaptation,” in *Proceedings of Interspeech*, 2023, pp. 1–5.

[2] H. Kargarandehkordi, M. Smith, and P. Fung, “Adaptive personalized speech emotion recognition,” in *Proceedings of ICASSP*, 2024, pp. 1–5.

[3] X. Li, J. Zhang, and H. Wang, “Speaker-dependent emotion recognition with personalization,” in *Proceedings of AAAI*, 2024, pp. 1–7.

[4] K. Ohene-Djan, N. Binetti, and R. Ang, “Emotional subtitles: A system and potential applications for deaf and hearing impaired people,” in *Proceedings of the ACM Multimedia*, 2006, pp. 1–4.

[5] Davoudi Mohsen, Mohammad Bagher Menhaj, Nima Shams Naraghi, Ahmad Aref, Mehran Davoodi, and Mohammad Davoudi, “A fuzzy logic-based video subtitle and caption coloring system,” *Advances in Fuzzy Systems*, vol. 2012, pp. Article ID 671851, 2012.

[6] Philipp Geuder, Marie Claire Leidinger, Martin von Lupin, Marian Dörk, and Tobias Schröder, “Emosaic: Visualizing affective content of text at varying granularity,” *arXiv preprint arXiv:2002.10096*, 2020.

[7] Y. Zhang et al., “Secap: Llama-based emotional captioning for subtitles,” in *Proceedings of ICASSP*, 2023, pp. 1–5.

[8] T. Ubur, L. Smith, and M. Cohen, “Augmented reality subtitles with multimodal emotion rendering,” in *Proceedings of the ACM Multimedia*, 2025, pp. 1–10.

[9] H. Wu, J. Lee, and S. Kim, “Emotion-aware subtitles for movies via speech analysis,” in *Proceedings of IEEE ICME*, 2025, pp. 1–6.

[10] B. M. Gorman, M. Crabb, and M. Armstrong, “Adaptive subtitles: Preferences and trade-offs in real-time media adaptation,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–11, ACM.

[11] Bruno Korbar, Jaesung Huh, and Andrew Zisserman, “Character-aware audio-visual subtitling,” in *Proceedings of ICASSP 2024*, 2024.

[12] D. Jonauskaite et al., “Cross-cultural color–emotion associations,” *Psychological Science*, vol. 31, no. 12, pp. 1513–1525, 2020.

[13] Anya C. Hurlbert and Yazhu Ling, “Biological components of sex differences in color preference,” *Current Biology*, vol. 17, no. 16, pp. R623–R625, 2007.

[14] Stephen E. Palmer and Karen B. Schloss, “An ecological valence theory of human color preference,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 107, no. 19, pp. 8877–8882, 2010.

[15] Abdulrahman S. Al-Rasheed, “Color preference theory: An ecological valence theory of human color preference in a saudi arabian sample,” *Psychological Reports*, vol. 116, no. 1, pp. 1–12, 2015.

[16] Abdulrahman S. Al-Rasheed, “Cross-cultural evidence for ecological valence theory of color preference,” *Frontiers in Psychology*, vol. 13, pp. 856112, 2022.

[17] Domicele Jonauskaite, Nele Dael, Laure Chèvre, Benno Althaus, Tiberiu Tremea, Louisa Charalambides, and Christine Mohr, “What is your red? investigating color–emotion associations in a large cross-cultural sample,” *Color Research & Application*, vol. 46, no. 5, pp. 1010–1026, 2021.

[18] Y. Wang et al., “Audio-only speech emotion recognition baseline,” in *Proceedings of ICASSP*, 2025.

[19] A. Kumar and S. Singh, “Gasmer: Graph-based speech emotion recognition,” in *Proceedings of ICASSP*, 2023.

[20] Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg, “Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances,” in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico, Apr. 2025, pp. 2202–2218, Association for Computational Linguistics.